Machine Learning

# Machine-learning data\_preprocessing

step1

import the required libraries (numpy, matplotlib.pyplot, pandas)

## step2

import the dataset

use pandas to import the dataset, save your .py file in the folder where the dataset is located

F5 will run the whole file and set the directory as working directory

X will contain all the independant data

[:, :-1].values = select all the rows and then select all the columns -1

(which is the purchased column we don't select)

y contains all the dependant data

[:, 3].values = select all from the 3rd column (3 index)

step3

take care of missing data

from sklearn.preprocessing import Imputer

imputer will replace the missing values (check imputer help CTRL + I)

imputer.fit() takes the matrix which contains the missing data

and selects on which columns the missing data are present

then we take this column with the index and apply imputer.transform() on it

## step4

encoding the cathegorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

the country and purchased column both contain the same names multiple times

(france, spain, germand, yes and no)

we have to encode this with the label and onehotencoder

step5

Splitting the dataset into the Training set and Test set

splitting the dataset in a training and test set are required to do

so the machine can learn from the training set and then test its finding with the test set

step6

we have to apply feature scaling so that the values aren't so far apart

when calculating the Euclidean distance

# Simple Linear Regression

**Formula =>**  y = b0 + b1 \* x1

y = dependent variable

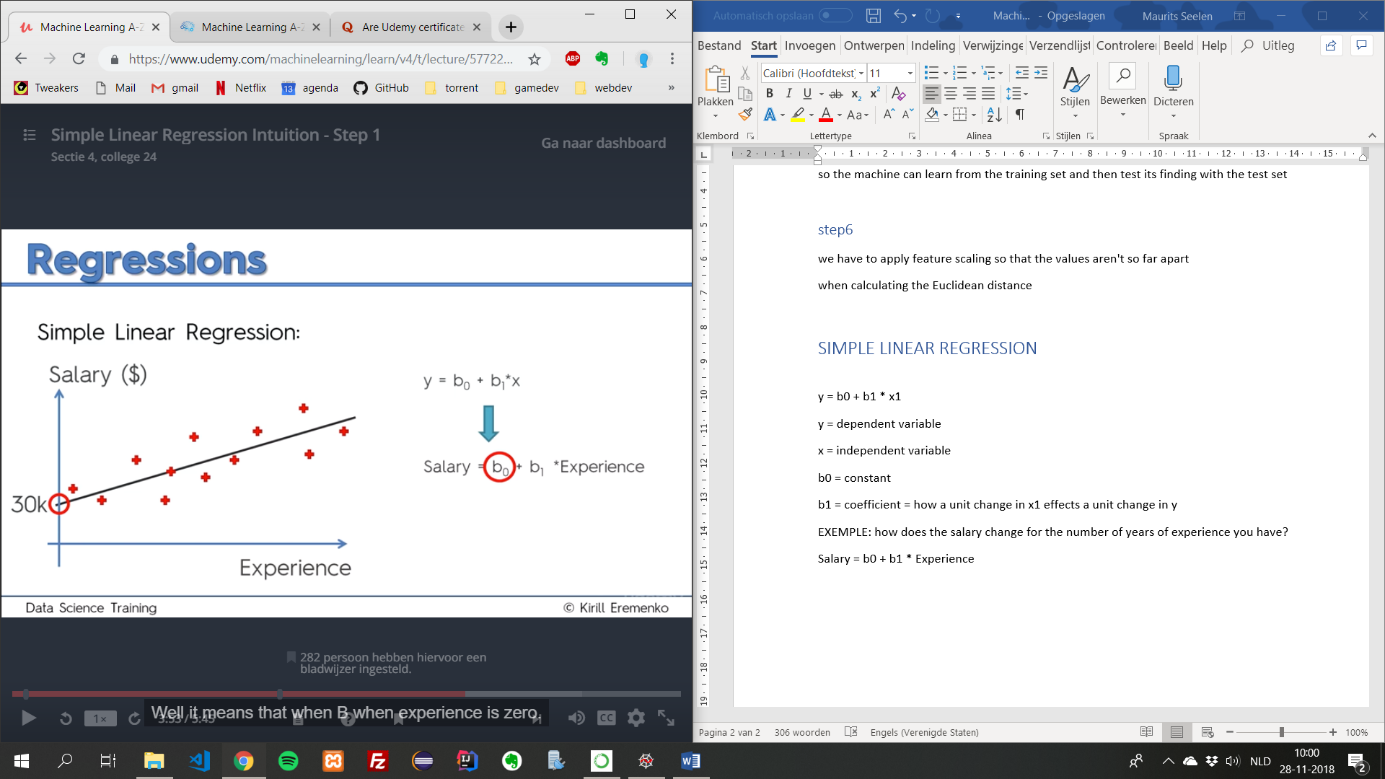
x = independent variable

b0 = constant

b1 = coefficient = how a unit change in x1 effects a unit change in y

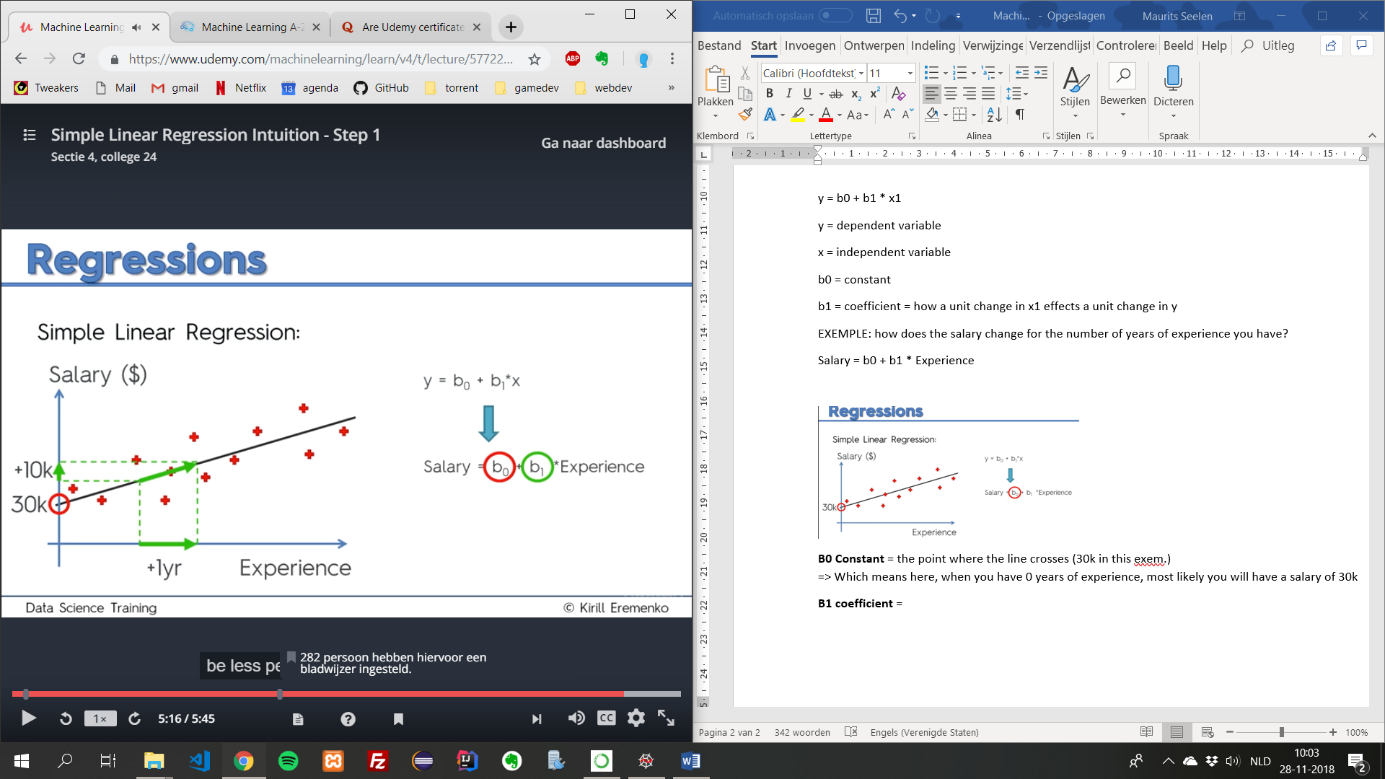
EXEMPLE: how does the salary change for the number of years of experience you have?

Salary = b0 + b1 \* Experience



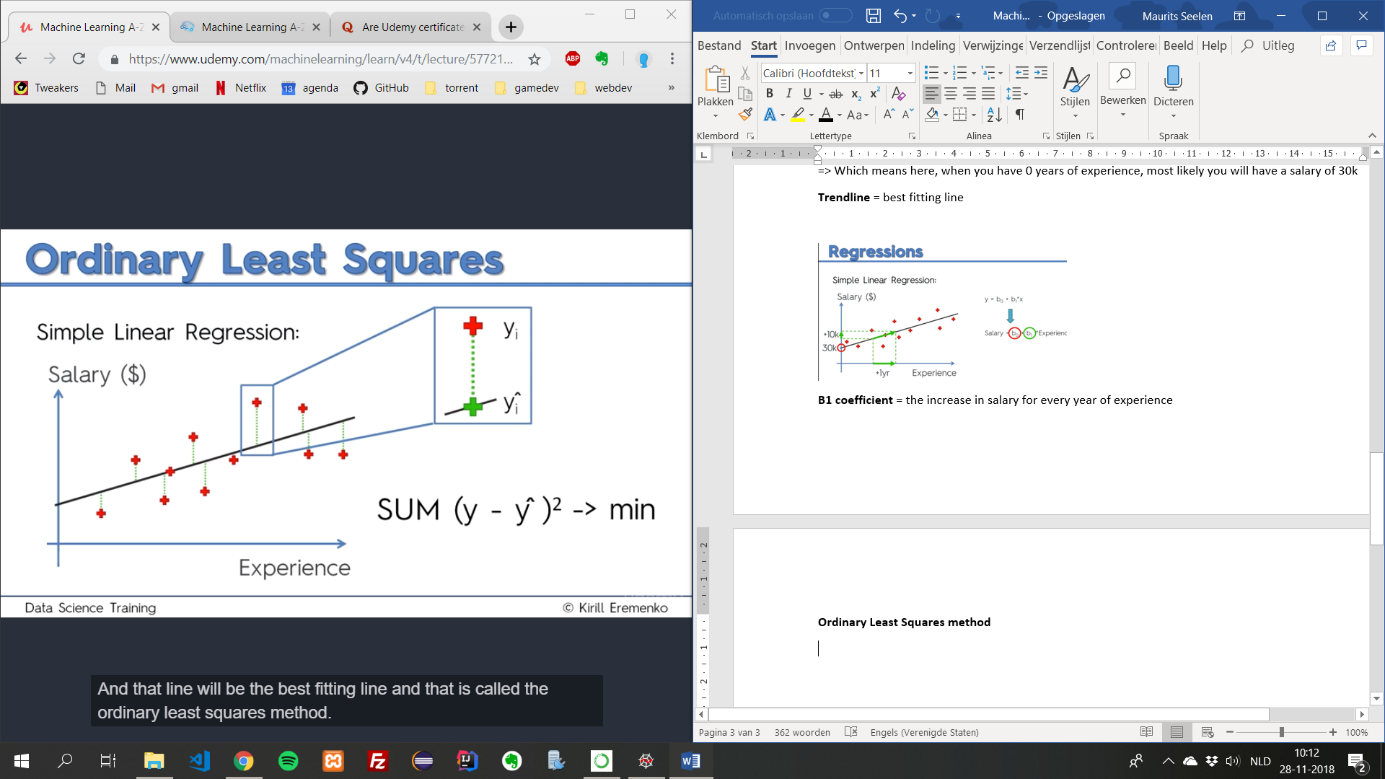
**B0 Constant** = the point where the line crosses (30k in this exem.)  
=> Which means here, when you have 0 years of experience, most likely you will have a salary of 30k

**Trendline** = best fitting line



**B1 coefficient** = the increase in salary for every year of experience

**Ordinary Least Squares method**



# Multiple Linear Regression

Y = b0 + b1\*x1 + b2\*x2 + …

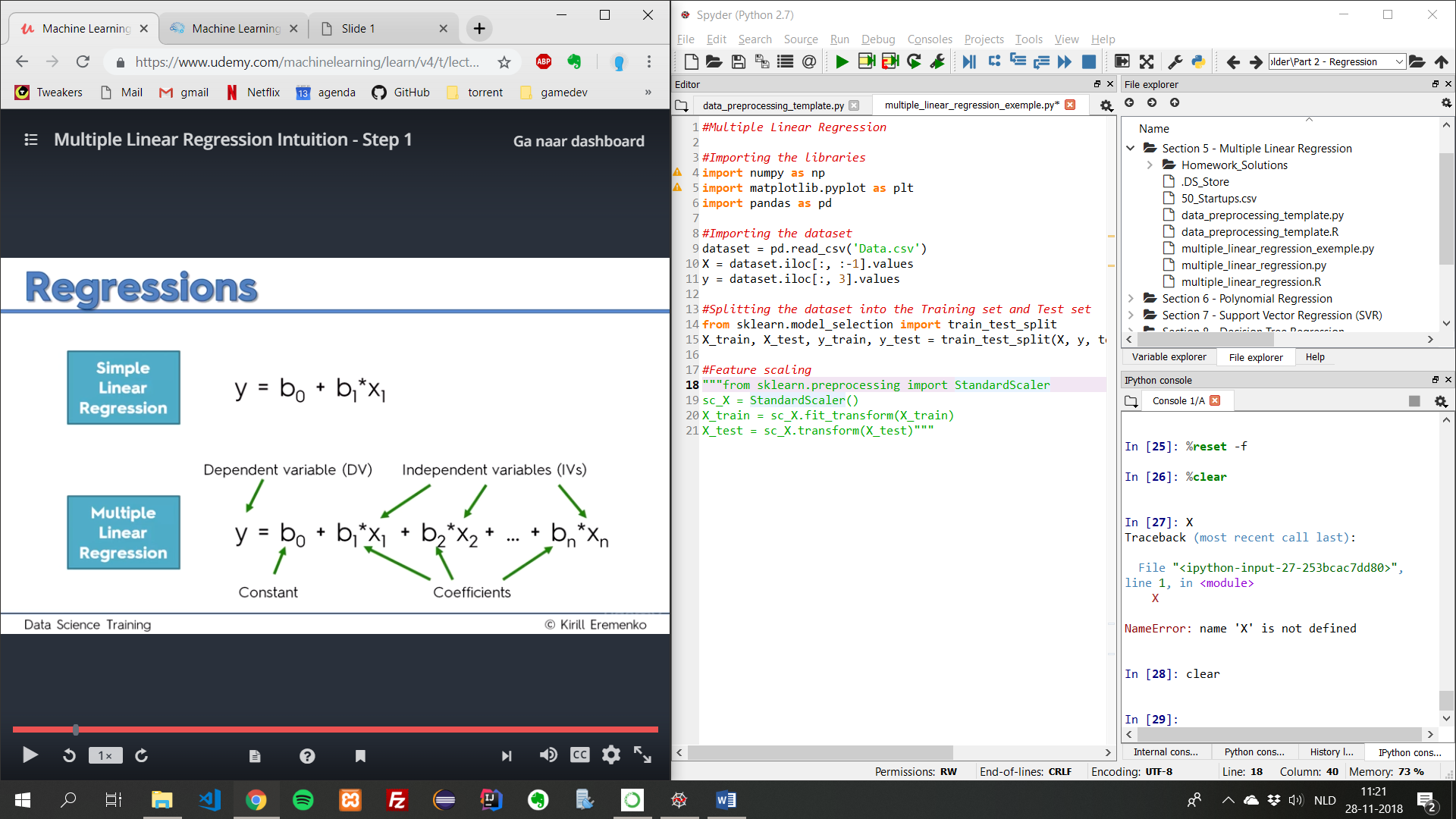
Y = dependent variable

X1, x2 = independent variables

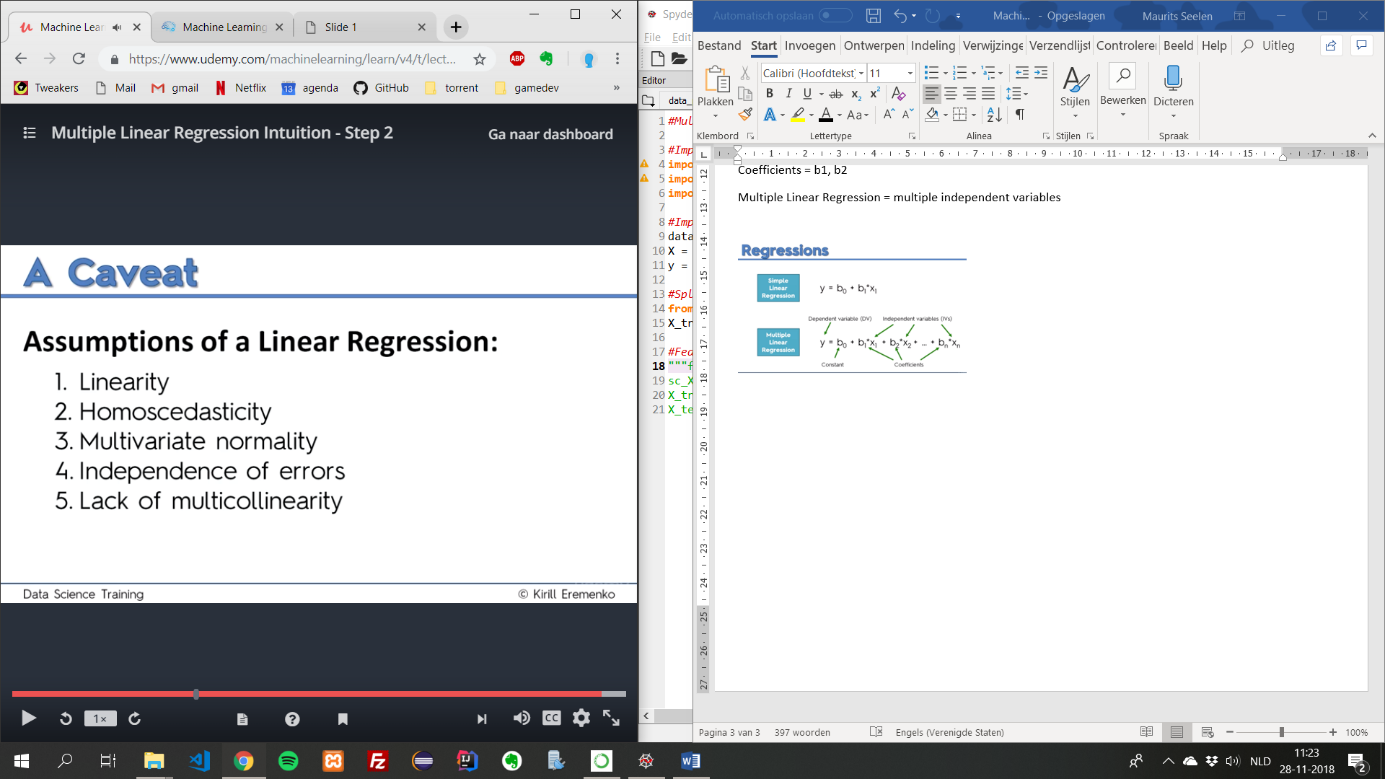
B0 = constant

Coefficients = b1, b2

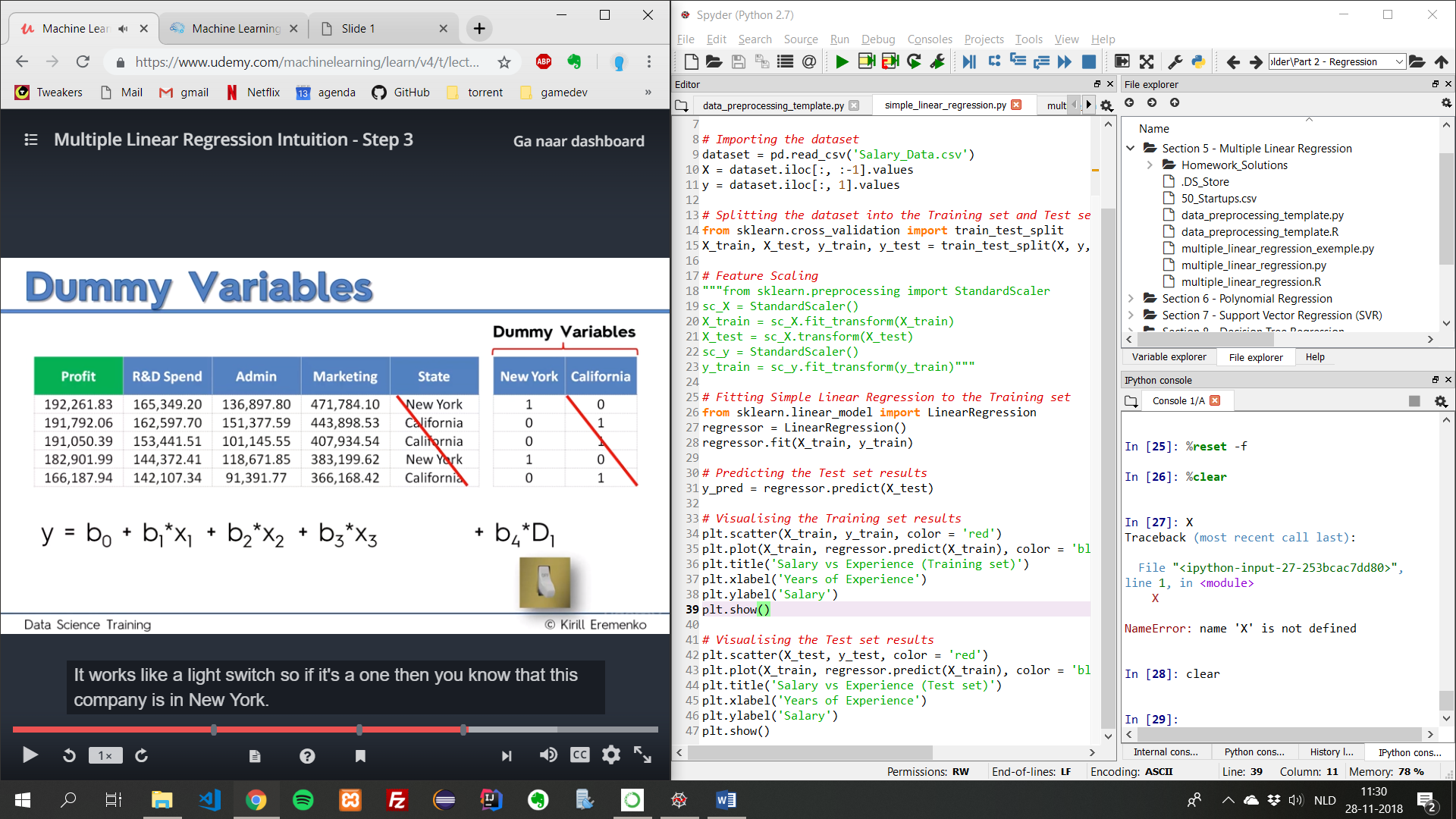
Multiple Linear Regression = multiple independent variables



You need to research if these assumptions are correct before building a Linear Regression Model.



## Dummy Variables



Always omit one dummy variable

If you have 10 dummy variables, only use 9, etc.

# Building a model

See pdf